What is claimed is:

1.   A computer-implemented method of identifying a candidate gene from a plurality of nucleotide sequences, the method comprising:

obtaining gene expression profile data for a plurality of nucleotide sequences, wherein said gene expression profile data describe behavioral patterns of gene expression;

identifying a group of said sequences for further analysis;

using information extraction algorithms to retrieve and extract pathway
information from a database comprising biological data;

cross-referencing said pathway information; and

viewing said cross-referenced information,

wherein viewing said cross-referenced information facilitates the identification of a candidate gene.

2.   The computer-implemented method of Claim 1, wherein said pathway information is stored in a database.

3.   The computer-implemented method of Claim 2, wherein said cross-referenced information is stored in a database.

4.   The computer-implemented method of Claim 1, wherein said cross-referenced information is viewable as a directed graph.

5.   The computer-implemented method of Claim 1, wherein identifying a group of sequences further comprises the step of clustering the gene expression profile data.

6. The computer-implemented method of Claim 5, wherein said clustering is unsupervised clustering.

7. The computer-implemented method of Claim 5, wherein said clustering is supervised clustering.

8. The computer-implemented method of Claim 5, wherein said clustering is a combination of supervised and unsupervised clustering.

9. The computer-implemented method of Claim 5, wherein said group of sequences represents a cluster.

10. The computer-implemented method of Claim 1, wherein said gene expression profile data is derived from microarray experiments.

11. The computer-implemented method of Claim 1, wherein said using information extraction algorithms is using natural language processing algorithms.

12. The computer-implemented method of Claim 11, wherein said natural language processing algorithms include template filling or Hidden Markov-Models.

13. The computer-implemented method of Claim 11, wherein said information extraction algorithm utilizes a text comparison algorithm.

14. The computer-implemented method of Claim 1, wherein said pathway information is extracted from one or more literature databases selected from the group consisting of MEDLINE, USPTO patent published patent database, USPTO issued patent database, the WIPO patent database, and the KEGG, MIPS and OMIM database.

15.   The computer-implemented method of Claim 14, further comprising the step of ranking the pathway information based on a ranking of a publication in a citation index.

16.   A data processing system for identifying candidate genes from a list of genes of known expression pattern, comprising:

a processor

a memory coupled to the processor, the memory configured to store instructions for execution by the processor, the instructions comprising:

instructions for accessing a list of genes of known expression pattern;

instructions for accessing and extracting pathway information from a literature database relevant to individual genes on the list of genes;

instructions for cross-referencing said pathway information; and

instructions for viewing said cross-referenced information.

17.   The data processing system of Claim 16, wherein said executable instructions further comprise instructions for storing said pathway and said cross-referenced information in a database.

18.   The data processing system of Claim 16, wherein said instructions for accessing the information sources comprise instructions for accessing a biomedical publication.

19.   The data processing system of Claim 17, wherein said executable instructions further comprise instructions for ranking said biomedical publication and instructions to assign a ranking score to said pathway information extracted from a biomedical publication based on the ranking of said biomedical publication.

20. A data processing system for identifying a candidate gene from a plurality of sequences, comprising:

a processor

a memory coupled to the processor, the memory configured to store instructions for execution by the processor, the instructions comprising:

instructions for clustering the plurality of sequences based on patterns of expression of the sequences, as described by gene expression profile data;

instructions for accessing and extracting information from a literature database;

instructions for cross-referencing said information; and

instructions for viewing said cross-referenced information.